Supplementary Information for 'Reflect – Augmented Browsing for the Life Scientist'

METHODS

User Interface. Both the browser plug-in and the web interface were constructed using HTML, JavaScript, XML-based User Interface Language, and Document Object Model events; communication between browser and server occurs via XMLHttpRequest objects.

Organism. By default, Reflect assumes protein names refer to human. The Firefox plug-in allows the user to change this default at any time to any of 373 organisms. In addition, the text of each HTML document is initially parsed to find recognized organism names, which are then added to the list of possible organisms for proteins in that document.

Tagging. After parsing for organism names, the text of the HTML document is parsed a second time for protein names. Parsing is done using leftmost longest matching of up to five words, testing each combination against the Reflect dictionary, which is stored in a hash table with all synonyms and orthographic variations occurring as hash keys. Recognized gene, protein, or small molecule names that occur in the text portion of the HMTL are then substituted with tags that call a Javascript function to generate the summary popups (Fig. 1). The document is then returned to the user's browser with previous HTML tags and their attributes unaffected, hence preserving the original document format.

Reflect Dictionary. The core component of Reflect is a consolidated dictionary that links synonyms to source data identifiers. The protein entries were derived from STRING[12], which in turn was created by importing the completed genomes in Ensembl, TAIR, Genome Review, and RefSeq (in this order of preference). In cases where one gene has several splice variants, the longest was chosen. In addition to importing all names and database accession numbers, we extended the dictionary with additional names from UniProtKB. The small molecule entries were derived from STITCH[8], which in turn was created by importing the compounds entries in PubChem[7]. Stereo-isomers were merged based on their canonical SMILES strings, and salt forms and trademark drug names were added as synonyms of the active substance. The dictionary is loaded into a Perl hash with each unique synonym (including all orthographic variations) as a unique key; this enables fast tagging, currently at a cost of 18 GB of RAM.

Popups. The summary popups are generated using overLib (http://www.bosrup.com/web/overlib/) and with content supplied mostly by CGIs on the Reflect server. For proteins and genes, the popup shows the same synonyms used by the Reflect dictionary, except that database identifiers and orthographic variations are not shown. The Reflect database identifier is then used to fetch the sequence and domain information from SMART[9], and to fetch the image of the five most significant interaction partners from STITCH[8]. Scrolling of the sequence and synonym lists was implemented in Javascript. The links to the best matching 3D structure in PDBsum[10] and the information about subcellular location were pre-calculated from the sequence database entries. The organism images were taken from iTol[13]. For small molecules, the popup shows the 2D structure from PubChem[7], and the five most significant interactions are derived from STITCH.

Accuracy. The accuracy of Reflect tagging was assessed using the BioCreative[11] benchmarks for *Saccharomyces cerevisiae* and *Drosophila melanogaster* (task 1B). For both organisms, we used Reflect to tag 250 short texts, and we used the BioCreative 'gold standard' genes to calculate an F-score (equal to the geometric mean of precision and recall). BioCreative also includes a mouse benchmark, however we could not use it due to difficulties with converting the gene identifiers into those required for Reflect.

Extending Reflect. We designed Reflect to be an extendible platform that facilitates adding further entity types. For each entity type, the Reflect dictionary needs a list that maps each synonym to an identifier; in addition, for each entity type, Reflect needs the address of a web service that, when combined with an identifier, can create the popup content for a single entity.

Evangelos Pafilis*, Seán I. O'Donoghue*, Lars J. Jensen*, Heiko Horn, Michael Kuhn, Nigel P. Brown, & Reinhard Schneider

European Molecular Biology Laboratory, 69117 Heidelberg, Germany.
e-mail: contact@reflect.ws
**These authors contributed equally.*

1. von Mering, C. et al. Nucleic Acids Res 35, D358-362 (2007).
2. Kuhn, M., von Mering, C., Campillos, M., Jensen, L.J. & Bork, P. Nucleic Acids Res 36, D684-688 (2008).
3. Hirschman, L., Colosimo, M., Morgan, A. & Yeh, A. BMC Bioinformatics 6 Suppl 1, S11 (2005).

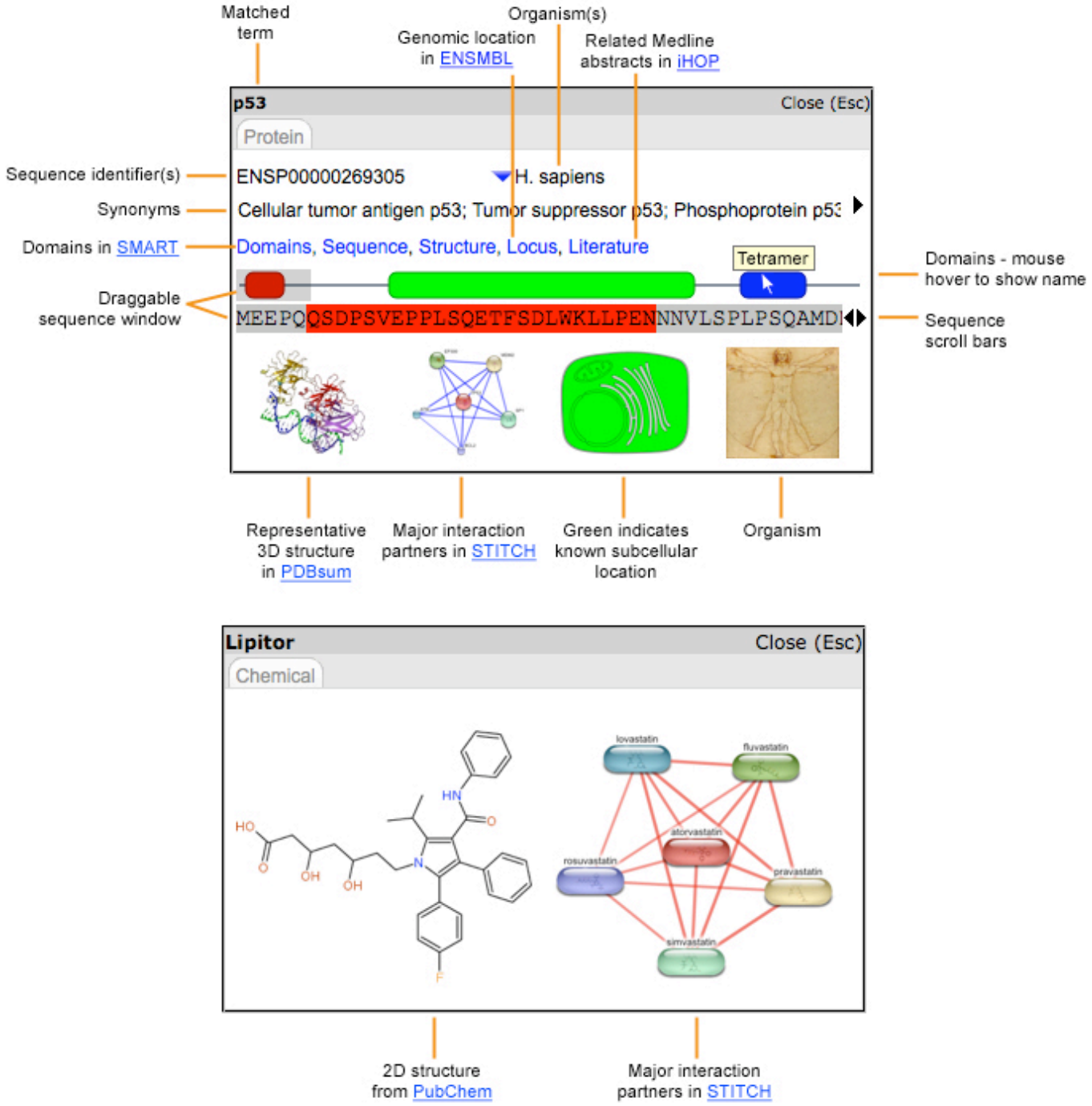

Figure 1 Details of summary popup content and links to more detailed information.